DOCUMENT RESUME

ED 281 873 TM 870 270

AUTHOR Gonzalez-Tamayo, Eulogio
TITLE Content Specifications of a Test and Generalizability Theory.
PUB DATE 87
NOTE 23p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adults; Analysis of Variance; *Content Validity; Criterion Referenced Tests; *Generalizability Theory; *Licensing Examinations (Professions); Multiple Choice Tests; Norm Referenced Tests; *Test Construction; Test Items; Test Reliability; *Test Theory; Weighted Scores
IDENTIFIERS *Internal Consistency; *Intraclass Correlation; Nested Data

ABSTRACT

The concepts of universe of admissible observation and universe of generalization from the generalizability theory were applied to calculate the intraclass correlation coefficient of a licensure test. The internal consistency coefficient of a dichotomously scored test is identical to the intraclass correlation coefficient of a two-facet design. The intraclass correlation coefficients for both norm- and criterion-referenced decisions were smaller in the nested than in the simple two-facet design and they were also smaller than the internal consistency coefficient. The proper design for most licensure tests and for tests using a design where the items are selected from content areas specified in the blueprint rather than from the universe of possible items must use a nested design to calculate the stability of test results rather than the internal consistency coefficient. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED281873

CONTENT SPECIFICATIONS OF A TEST AND GENERALIZABILITY THEORY

Eulogio Gonzalez-Tamayo

Board of Examiners

New York City Board of Education

1987

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

E. Gonzalez-Tamayo

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 870 270

Running Heading: CONTENT SPECIFICATIONS

BEST COPY AVAILABLE

SUMMARY

The concepts of universe of admissible observation and universe of generalization from the generalizability theory were applied to calculate the intraclass correlation coefficient of a licensure test. The internal consistency coefficient of a dichotomously scored test is identical to the intraclass correlation coefficient of a two-facet design. The intraclass correlation coefficients for both norm- and criterion-referenced decisions were smaller in the nested than in the simple two-facet design and they were also smaller than the internal consistency coefficient. The proper design for most licensure tests and for tests using a design where the items are selected from content areas specified in the blue-print rather than from the universe of possible items must use a nested design to calculate the stability of test results rather than the internal consistency coefficient.

INTRODUCTION

Among the descriptive interpretations of a test, one important consideration is its cotent validity. Test developers, generally, describe one procedure of selecting the test questions and point to a broader universe that the test supposely represents. A necessary but insufficient condition for the test to be content valid is that the questions be part of the specifications or that they are related to the content. Content validity is more than just content relatedness and it also involves (1) sampling from a specified universe and (2) allocating weights to each subpopulation of essential characteristics on some scientific bases.

The concept of sampling of items or observations from a universe has been criticized as impractical (Thorndike, 1966) and even improper (Loevinger, 1965). In licensure examinations content validity is usually the only validity available and eliminating the concept of sampling would reduce the test to a series of exercises that are job related but with very limited meaning or usefulness. Present methods of determining content validity are safe and cheap to implement but they are inadequate and designed to confirm the hypothesis of relatedness (Madaus, 1986).

Content validity is related, statistically, to the stability of a selection procedure. The items included in the test represent a sample of the universe of items and it is a relevant questions to investigate the stability of the results of the selection procedure regardless of the validity of the test. Measurement issues in psychology have, historically, been addressed through the classical test theory that postulates that an observed measurement can be decomposed into true score and a single source of random error. As such, the classical test theory model can not clearly differentiate among multiple sources of error. Scores in a test where the items have been selected from different subdomains present several sources of variability. These multiple sources of variability, unmanageable in the classical theory of tests, con be studied and analyzed through the concepts of the modern theory of generalizability.

Fundamental Concepts of Generalizability Theory

The generalizability theory, Cronbach et Al.(1972), represents the most comprehensive approach to the study of reliability in general. The theory emphasizes the estimation, use, and interpretation of variance components associated with universes of admissible observations (Brennan, 1983).

Universe of Admissible Observation (G Study)

In developing a licensure examination one may decide to select a variety of items representing the knowledges, skills and abilities of the content covered in the training required from those applying for the license or for the successful performance of the duties of the license. Any one of the possible items constitutes an admissible and acceptable condition of measurement for the item facet; similarly, any of the possible examinees constitutes an admissible and acceptable condition of measurement for the person facet.

Given this universe of admissible observations, it is necessary to collect and analyze data to characterize empirically the universe of admissible observations. In a two-facet design there are three sources of variation in the observations that can independently be estimated: persons (p) and items (i) effects and the interaction between persons and items (p.i). All the items do not represent the same level of difficulty and there are variations in the responses from the examinees regardless of the person who is answering the items; all persons have greater probability of answering correctly an easy than a difficult item. The examinees do not possess the same level of ability and there is variation between the examinees' responses regardless of the item. Each examinee does not use the same level of ability in answering all the items and examinees show preference for some items and indifference for others regardless of the items difficulty and there is interaction between examinees and items.

In an empirical case, only a sample of examinees and a sample of items is used;

in another situation, different examinees and different items could be selected under the same specifications and different variance components would be obtained. Unbiased estimates of these variance components are easily obtained from the usual mean squares of an analysis of variance.

Universe of Generalization (D Study)

The purpose of a G study is to obtain the estimates of the variance components associated with the universe of admissible observations. A D study emphasizes the estimation, use, and interpretation of variance components for decision making with well specified procedures. There may be a need to make a decision about how to rank the examinees on the basis of their test scores or the decision may be to license only those reaching a predetermined minimum score. If the purpose is to rank the examinees, the systematic variance between items, the items are easier or more difficult than other possible items for all the examinees, would be irrelevant for the decision because all the examinees answered the same items. Another sample of items will show different variance but the ranking of the examinees will be the same and item variability does not affect the decision.

If, on the other hand, the decision is to license those examinees reaching a predetermined level of achievement in the test, then variability among items will be important in the final decision of which applicants are licensed and which are not. The fate of an applicant may depend on how easy or difficult were few items selected at random. The same examinee taking another test developed under the same specifications and where the items are also randomly selected from the same universe of items, may score differently. The fate of an applicant, however, should not depend on the characteristics of few items and item variability should be part of the D study consideration.

Simple and Nested Designs

A two-facet design, items and persons selected randomly, is the simplest design for a test but it is an unrealistic design in most testing situations and expe-

cially in licensure tests. The two-facet design implies that the items have been randomly selected from the universe of admissible observations or items. In licensure tests, items are selected at random within subdomains describing the test rather than from the universe of items.

The design for a licensure test may look like a three-facet desig with persons, subdomains, and items as the three dimensions. Most of these tests, however, are best viewed as being made up of items (i) nested within content categories (h) (Brennan, 1983). In such cases, it is reasonable to view the universe of admissible observations as a i:h design with every item in the universe being associated with one and only one content category (h). For such a design, a three-facet design (p.i.h) is a logical impossibility because each item is not associated with each content category in the universe and a nested design, p.(i:h) is the most powerful design.

There are three important issues that must be taken into consideration in a nested design. First, all content categories are usually represented in the design specifications and there is no selection of categories in an empirical case. Second, in most empirical cases, the number of items in each category is not the same; content categories are usually weighted and the common procedure to allocate weights is by the number of items included in each category. Third, the number of categories is usually small.

AN EMPIRICAL CASE

As part of the examination for the license as Teacher of Accounting and Business Practices, a group of 94 applicants answered a multiple choice type test. The test design, based on the job analysis, called for five different areas with different weights and different number of items in each component as reported in table 1. One

Table 1 about here

of the original items was rejected in the final analysis because of ambiguity in the question and the final number of valid items was 64 instead of the original 65. The designed contained 94 persons (p) taking a test that included five different areas (h) with unequal number of items in each area (i).

The raw data from the study are reported in table 2, and the meaning of the

Table 2 about here

symbols is as follows

$\sum X_{p.i.h}$ and $\sum(X_{p.i.h})^2$ indicate the sum of individual scores on each of the items (either zero or one) and the sum of squares of the individual scores on each item. These two sums are identical in a dichotomously scored test.

$(\sum X_{p.i.h})^2$ is the square of the sum of scores over all the persons and over all the items.

$\sum(X_{.i.h})^2$ represents the sum of squares of scores across persons for each item and there are 64 terms in the sum, one for each item.

$\sum(X_{p..})^2$ represents the sum of squares of scores across items and components, or the sum of squares of individual scores in the test, and there are 94 terms in the sum, one for each of the applicants.

$\sum(X_{p.1})^2$ represents the sum of squares of individual scores on component one; there are five sums like this in the study, one for each of the components, and each sum contains 94 terms, one for each applicant.

$(X_{..1})^2$, without the summation sign, represents the squares of the sums of scores in each component across all the persons and all the items in the component; there are five of these sums.

The data presented in table 2 can be analyzed in two different G designs and for two different D decisions. The test can be considered as a random sample of 64 items from the universe of items, a simple two-facet design of the form p.j, or the test may be viewed as five random samples of items within each of the five areas of the test and the design would be a nested design of the form p.(i:h). The data from the analysis of variance, G study, can be used for two different D decisions: for norm- and for criterion-related decisions.

A Two-Facet Design

In a two-facet design of the form p.j, there are 94 persons (p) and 64 items (i). In the design, it is assumed that the items were selected randomly from the universe of items rather than from each of the areas of the test.

Genova Analysis

The elements needed to calculate the different components of the analysis of variance are calculated from the data of table 2 and they are

$$\sum(X_{p.j}) = \sum(X_{p.j})^2 = 3702$$
$$(\sum X_{p.j})^2/(p.j) = 2278.0592$$
$$\sum(X_{p.})^2/j = 2365.7188$$
$$\sum(X_{.j})^2/p = 2597.5106$$

The different components of the variance are

$$S_T = 3702 - 2278.0592 = 1423.9408$$

$$S_p = 2365.7188 - 2278.0592 = 87.6596$$

$$S_j = 2597.5106 - 2278.0592 = 319.4514$$

$$S_{pj} = S_T - S_p - S_j = 1016.8298$$

The complete analysis of variance is reported in table 3, where,in addition to

Table 3 about here

the sum of squares, the degrees of freedom and the respective mean squares are also included.

The Intraclass Correlation

The intraclass correlation coefficient can easily be calculated in a two-facet design from the mean squares without the need for the estimates of the true variance. The interpretation and use of the components of the variance will depend on the kind of decisions expected or on the D design.

If the interest is to rank applicants according to their total score, norm-referenced decisions, there are only two relevant sources of variation: person variability and interaction variability. Item variability will not affect the ranking for the simple reason that all the applicants answered the same items and any constant bias in the items, the items in the sample are easier or more difficult than the items in any other possible sample of items, will affect all scores in the same way; the ranking of applicants will remain unchanged from one sample of items to another possible sample of items. Persons variability represents the observed variance; interaction variability represents the error variance. The intraclass correlation coefficient, defined as the ration of true to observed variance and considering that the true variance is equal to the observed minus the error variance, will be given by

$$R_{norm} = (0.9426 - 0.1736)/0.9426 = 0.8159$$

If the decision requires a comparison of the individual scores with a criterion,

criterion-related decisions, then variability among the items reflects error impacting on the decision. In a licensure examination, the scores are compared with a predetermined criterion or passing score; those applicants reaching the criterion are licensed and those failing to reach the criterion are not. An individual score, and the final fate of an applicant of being licensed or not, is affected by how easy or difficult the items included in the test are and item variability from one testing situation to another can make the difference between reaching or failing to reach the criterion. In such a case, item variability is part of the error and the intraclass correlation coefficient for criterion-referenced testing is given by

$$R_{criterion} = \frac{0.9426 - 0.1736}{0.9426 + (5.0707 - 0.1736)/64} = 0.7546$$

A Nested Two-Facet Design

In the present study, the proper analysis of variance design is a nested design of the form p.(i:h) where the items are selected at random within each of the five components of the test. Regardless of how the items are selected, the test must always contain 25 items dealing with Bookkeeping, 14 items dealing with Business Math, and so on with each of the five components.

Genova Analysis

There are five sources of variance in the present nested design: persons variability (p), component variability (h), item variability within the component (i:h), interaction between persons and components (p.h), and interaction between persons and items within components, p.(i:h). The interaction between items and components and the triple interaction between persons, items and components are eliminated in the nested design; the items in each component are different items and it is not possible to separate the interaction of items and components from the error.

The total sum of squares, the sum of squares for persons and the sum of squares

for items are the same in the simple than in the nested design. The total number of degrees of freedom and the number of degrees of freedom for persons are also the same in both designs. The items in the nested design, however, can only variate within each component and the degrees of freedom are the sum of items in each component minus one or the number of items minus the number of components (64 - 5 = 59).

The sum of squares for components variability are $\sum(X_{p.i})^2, \ldots\ldots \sum(X_{p.5})^2$ and these sums have to be divided by the number of terms which is different in each of the five sums. These divisors are 94x25 for component one, 94x14 for component two, 94x10 for components three and four, and 94x5 for component five, or the number of persons times the number of items in each component. The variance component for item variability will be

$$(204204)/(94x25) + (710648)/(94x14) + (331776)/(94x10)$$
$$+ (32604)/(94x10) + (80089)/(94x5) - 2278.0592 = 1.1086$$

and the number of degrees of freedom will be the number of components minus one (4).

The sum of squares for the interaction between persons and components are calculated from the symbols $(X_{..1})^2, \ldots.(X_{..5})^2$ of table 2. The divisor for each square is given by the number of items in the component; there are as many interactions of persons within each component as there are items in the component. The sum will be

$$(23095)/25 + (8115)/14 + (3964)/10$$
$$+ (3729)/10 + (931)/5 - 2278.0592 = 180.8837$$

and the number of degrees of freedom for the total sum will be the number of persons minus one times the number of components minus one or $(p-1)(h-1) = 372$.

Finally, the sum of squares for the interaction between persons and items within the components ,p.(i;h) will be the total sum of squares minus the sum of squares for the components already calculated

$$S_{p.(i:h)} = S_T - S_p - S_{i:h} - S_h - S_{p.h} = 834.8375$$

and the number of degrees of freedom will be the total number of degrees of freedom (6015) minus the nymber of degrees of freedom already allocated or 5487.

The mean squares are easilly calculated by dividing the sum of squares by the

respectives degrees of freedom as reported in table 4. The estimates of the true

Table 4 about here

variance components, fundamental for the calculation of the intraclass correlation coefficients in the nested design, are reported in table 5, and they are based

Table 5 about here

on the formulas proposed by Brennan (1983) for a nested design. A modification was necessary for the effect of items within components (i:h), because each component has different number of items; the factor selected was the average number of items per component or 64/5 = 12.8. This number is an approximation, but so are the other estimates of the true variance components.

Intraclass Correlations

In the nested design there is no variability in the components from one testing situation to another because any test developed under the same specifications must include the same five components. The differences between two testing situations would be in the items included in each component and the variability would be in the items but not in the components. The components variance and the interaction variance between persons and components are not considered in the calculation of the intraclass correlation.

The difference between a norm- and a criterion-referenced decision will be given by the variation due to the items within the components or $\hat{\sigma}^2_{i:h}$. This component of the variance is irrelevant in a norm-referenced decision because all the applicants will answer the same items and any systematic bias in the items will affect equally all the applicants but not their rankings. This variance component is important in a criterion-related situation; the fact that the items in a particular test are easier or more difficult than in other possible test developed under the same specifications, could make the difference between reaching or failing to reach a minimum qualifying

score. The interaction effect between persons and test components is not error because the same test components will be included in any testing situation and, therefore, there is no variation in test components from one to another testing situation. The interaction between persons and items within the components is, however, an error and it could be different from test to test under the same specifications.

The intraclass correlation coefficient for norm-referenced decisions would be

$$R_{norm} = \frac{0.0071}{0.0071 + (0.1521)/12.8} = 0.3737$$

and the coefficient for criterion-referenced decisions would be

$$R_{criterion} = \frac{0.0071}{0.0071 + (0.1521 + 0.0560)/12.8} = 0.3032$$

Discussion of the Results

The intraclass correlation coefficients calculated from a Genova analysis for a simple and for a nested two-facet design and for different kinds of decisions (D studies) are reported in table 6. There are substantial differences in the values

---

Table 6 about here

---

of the intraclass correlation coefficients depending on the G design and also on the D study or the kind of decisions for which the test results will be used.

The coefficients were lower for criterion- than for norm-referenced decisions in both the simple and the nested two-facet designs. In any test, the reliability coefficients are never higher for criterion than for norm-referenced decisions because the variability between items does not affect norm-referenced results but it impacts upon the criterion-referenced results.

There was a very large difference in the value of the coefficients two G designs. The intraclass correlation coefficients, for both norm- and criterion-

referenced decisions, are much larger in a simple two-facet than in a nested design. The coefficients were more than twice larger in a simple two-facet design (0.82 and 0.75 for norm- and for criterion-referenced decisions respectively) than in the nested design (0.37 and 0.30 respectively). As a matter of fact, the coefficients were so low for the nested design that they suggest the inadvisability of using the test results for any kind of decision, either norm- or criterion-related decisions.

The internal consistency of a test, or Kuder-Richardson formula 20 for items scored dichotomously, is always equal to the intraclass correlation coefficient of a simple two-facet design. In the present empirical case, the internal consistency was KR_{20} = 0.8180 and the intraclass correlation coefficient in the simple two-facet design for norm-referenced decisions was R_{norm} = 0.8169 and the two formulas provide practically the same coefficient. This result is not an artifact of this data. For a simple two-facet design and for norm-referenced decisions, the intraclass correlation coefficient is identical to the KR_{20} when items are scored dichotomously and for other metrics, the intraclass correlation coefficient is identical to the coefficient alpha (Brennan, 1983). A simple arithmetic transformation of the elements of one of the formulas to the elements of the other would demonstrate the identity of both formulas.

The intraclass correlation coefficient from a simple two-facet design is never smaller than the intraclass correlation coefficient from a two-facet nested design. The two designs have the same number of degrees of freedom and the same sums of squares for total variability and for persons and item variability. The remaining of the total variance is a single source of variability in the simple two-facet design, interaction between persons and items, but it is decomposed into three sources of variation in the nested design: component variability, interaction between persons and components, and interaction between persons and items. The reamining degrees of freedom are larger in the simple than in the nested design. The sum of squares, after eliminating the sum of squares for persons and items, would be represented in the following symbolic form for the two designs

$$\sum x^2 = \sum x_1^2 + \sum x_2^2 + \sum x_3^2$$

where the first part of the equation represents the only source of variation in the simple two-facet design and the second part of the equation represents the three sources of variation in the nested design. The degrees of freedom would be

$$N = n_1 + n_2 + n_3$$

and the N is always, without exception, larger than any of the other n's. The mean squares, therefore, will form the following inequality

$$(\sum X^2)/N < (\sum X_1^2)/n_1 + (\sum X_2^2)/n_2 + (\sum X_3^2)/n_3$$

and it is obvious that the first part of the inequality is smaller than the second part.

The mean squares that have been analyzed can be considered as error since any variability with the exception of true person variability contributes to the instability of the test results. The errors, therefore, are larger in the nested than in the simple two-facet design and, as a result, the intraclass correlation coefficient is smaller in the simple two-facet design than in the nested design and it is also smaller than the internal consistency coefficient as measured by the KR_{20}.

CONCLUSION

Achievement, aptitude, licensing, and most other psychological tests are developed in such a way that the proper design to calculate the intraclass correlation coefficient is the nested design. Items are rarely, if ever, selected randomly from the universe of all the possible manifestations of the attribute. The usual selection procedure is that the domain of the test is subdivided into subdomains and each subdomain is weighted differently and the items are selected rondomly within each subdomain or test component. The test blueprint tends to become most explicit and most full elaborated in the case of achievement testing in some academic field. It is common to prepare a two-dimensional grid showing in one dimension the areas to be covered by the test and in the other the types of operations to apply to each area of the content. In licensing and job proficiency testing, the test blueprint is in many ways similar to that of an academic achievement test. The emphasis, however, will be in the knowledges, skills and abilities necessary for the successful performance on the job. Content specifications for a test of aptitudes tend to be somewhat less full than specifications for achievement or proficiency testing; for tests of interest, temperament, and attitudes, content specifications are more sketchy with an elaborate definition of the trait trying to indicate some of the aspects of the content. The fact that some of the test specifications are more detailed than others is not because specifications are not needed in some of the tests but rather because test developers are unable to produce a detailed blueprint. Every test developer tries to produce a test that fairly represents all the important areas where the attributes can be manifested.

It does not make any sense to develop the test blueprint similar to a two-dimensional grid and then calculate the intraclass correlation coefficient, as an indication of the stability of the test results, on the bases of an analysis of variance performed as a simple two-facet design. The only time that a two-facet design could be used correctly to calculate the intraclass correlation coefficient would be

if many items, many times more than the items needed for the test, have been developed for each area of the test blueprint and this bank of items is considered as the universe of all possible manifestations of the attribute and then a simple random selection is made from this universe of items. As long as the items are selected from each area of the test blueprint the proper procedure to calculate the intraclass correlation coefficient is the nested design.

In many licensing and achievement tests, the only reliability coefficient that very often is calculated is the internal consistency coefficient. In licensing and job proficiency testing, different tests are used for every testing situation and the most easily available method to calculate the test reliability is through the internal consistency method. Internal consistency and consistency of the decisions based on test results are not the same thing. Popham and Husek (1969) argued that, since criterion-referenced testing are designed to determine a person's achievement compared to some criterion, the meaning of the score should not depend on the scores of other people and variability is not a necessary condition for a good criterion-referenced test and reliability indices based on scores variability are not only irrelevant to criterion referenced tests but may actually be injurious to their proper development and use. Kane (1986), however, presented an analysis suggesting that reliability is an important issue in criterion-referenced testing; in particular, if a criterion-referenced test had a reliability (defined in terms of intraclass correlation coefficient) below 0.50, a simple a priori procedure would provide better estimates of students universe scores than would individual observed scores.

The internal consistency coefficient is an estimate of the stability of the responses through all the items of the test. The calculation of the coefficient is based on the assumption that all items are measures of the same underlying attribute, that is, the test is homegeneous in content. In most licensing and job proficiency tests the assumption of homogeneity is not tenable; these tests are developed to include several factors, skills, abilities, or areas of knowledge. In licensing and job proficiency testing, the internal consistency coefficient is the wrong estimate of the

reliability of the test and it provides a higher estimate than the coefficient based on an analysis of variance.

The two-dimensional grid of the specifications for the development of a test, or a blueprint indicating the different components of a test in a job analysis, may be viewed as an statistical sophistication in test development. Specification of the areas of content and the functions or levels of operations included in each area plus the weights assigned to each cell of the grid is, no doubt, the best test design that psychometricians have been able to produce. The procedure, however, may be the best design to develop a large pool of items that will constitute the universe of all possible items from which the actual items to be included in the test will be selected; but it is not the best design to select the items for a test in terms of the generalizability of the results from the test. A simple random selection from the universe of items will produce a higher consistency of the results as indicated by the intraclass correlation coefficient than the nested design which will be the proper design for a two-dimensional grid of test development.

BIBLIOGRAPHY

Brennan, R. L. (1983). Elements of Genralizability Theory. The American College Testing Program: Iowa City, Iowa.

Cronbach, L.J., Gleser, G.C., Nanda, H., and Rajaratnam, N. (1972). The Dependability of Behavioral Measurements: Theory of Genralizability for Scores and Profiles. New York: Wiley.

Kane, M.T. (1986). The Role of Reliability in Criterion-referenced Tests. Journal of Educational Measurement, 23, 221-224.

Loevinger, J. (1965). Person and Population as Psychometric Concepts. Psychological Review, 72, 143-155.

Madaus, G.F. (1986) Measurement Specialists: Testing the Faith. A Reply to Mehrens. Educational Measurement Issues and Practice, 5, 11-14.

Popham, W.J. and Husek, T.R. (1969) Implications of Criterion-referenced Measurement. Journal of Educational Measurement, 6, 1-9.

Thorndike, R.L. (1966) Reliability. In A. Anastasi (Ed.), Testing Problems in Perspectiv Washington: American Council on Education.

Table 1: Components and Number of Items per Component in the Test for Teacher of Accounting and Business Practice

Component	Number of Items
1. Bookkeeping Theory	25
2. Business Mathematics	14
3. Business Law	10
4. Computers in Business	10
5. Teaching Methodology	5

Table 2: Raw Data for the Analysis of Variance in the Test For Teacher of Accounting and Business Practice

Symbol	Value	Symbol	Value
$\sum X_{p.i.h}$	3702	$\sum (X_{p..4})^2$	326041
$\sum (X_{p.i.h})^2$	3702	$\sum (X_{p..5})^2$	80089
$(\sum X_{p.i.h})^2$	13704804	$(X_{..1})^2$	23095
$\sum (X_{.i:h})^2$	244166	$(X_{..2})^2$	8115
$\sum (X_{p..})^2$	151406	$(X_{..3})^2$	3964
$\sum (X_{p.1})^2$	2042041	$(X_{..4})^2$	3729
$\sum (X_{p..2})^2$	710648	$(X_{..5})^2$	931
$\sum (X_{p..3})^2$	331776		

Table 3: Variance Components for the Design Persons and Items (p.j) in the Test for Teacher of Accounting and Business Practice

	Variance Data		
Source	Sum of Squares	d.f.	Mean Squares
Total	1423.9408	6015	
Persons	87.6596	93	0.9426
Items	319.4514	63	5.0707
p.j	1016.8298	5859	0.1736

Table 4: Variance Components for the Nested Design in the Test for Teacher of Accounting and Business Practice

	Variance Data		
Source	Sum of Squares	d.f.	Mean Square
Total (S_T)	1423.9408	6015	
Person (S_p)	87.6596	93	0.9426
Items ($S_{i:h}$)	319.4514	59	5.4144
Components (S_h)	1.1086	4	0.2772
p.h ($S_{p.h}$)	180.8837	372	0.4862
p.i ($S_{p.i:h}$)	834.8375	5487	0.1521

Table 5: Estimates of the True Variance for the Different Components of the Nested Design in the Test for Teacher of Accounting and Business Practice

Effect	Factor	Formulas	$\hat{\sigma}$
p	p = 94	$(MS_p - MS_{p.h})/(h.i)$	0.0071
h	h= 5	$(MS_h - MS_{i:h} - MS_{p.h} + MS_{p.i:h})/p.i$	- 0.0045(1)
i:h	*i = 12.8	$(MS_{i:h} - MS_{p.i:h})/p$	0.0560
p.h	---	$(MS_{p.h} - MS_{p.i:h})/i$	0.0261
p.i:h	---	$(MS_{p.i:h}$	0.1521

* This number is an average of the number of items per component

(1) The variance components can not be negative. Because they are estimates, sometimes the actual value of zero is estimated as negative but the value used in the formulas has to be zero.

Table 6: The Intraclass Correlation Coefficients for Two Genova Designs and Two D Studies in the Test for Teacher of Accounting and Business Practice

	D Study	
Genova Design	Norm-Reference	Criterion-Referenced
Simple Two-Facet (p.j)	0.8159	0.7546
Nested Design (p.i:h)	0.3737	0.3034